ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Title goes here

Master Thesis

Robin Bosshard, 16-915-399

October 16, 2023

Supervisors: Prof. Dr. Fernando Perez-Cruz, Dr. Eliza Harris, Lili Gasser (SDSC)
Dr. Kasia Arturi (Eawag)

Department of Computer Science, ETH Zürich

Abstract

This thesis enhances the MLinvitroTox framework, which predicts the toxicity of unknown compounds from HRMS/MS data. This framework can forecast the most hazardous compounds in environmental samples, circumventing the need for resource-intensive chemical identification. It uses machine learning models trained on ToxCast/Tox21 in vitro data and SIRIUS molecular fingerprints. We implemented a new processing pipeline called pytcpl and expanded its application to the latest toxicity data, achieving an average balanced accuracy of todo:X for binarized toxicity prediction. MLinvitroTox was also effective when validated with MassBank spectra data. Additionally, we developed a web app for user-friendly interaction with the framework.

Acknowledgments

First and foremost, I would like to thank Prof. Dr. Fernando Perez-Cruz from the Swiss Data Science Center (SDSC) for granting me the opportunity to work on this fascinating project. His support has been invaluable.

I would like to express my sincere gratitude to my supervisor Dr. Kasia Arturi from Swiss Federal Institute of Aquatic Science and Technology (Eawag) and my supervisors Dr. Eliza Harris, Lili Gasser from SDSC for their numerous discussions, patience and valuable insights. Without their help, this thesis would not have been achievable.

Additionally, I would like to acknowledge Prof. Dr. Juliane Hollender from Eawag for her support throughout the project and for the enlightening experience of visiting the Eawag labs.

Furthermore, my gratitude goes out to Jason Brown, Feshuk Madison, and Katie Paul Friedman for their participation in discussions concerning the technical aspects of the tcpl pipeline and the ToxCast database.

Lastly, I extend a special thank you to my family and friends for their unconditional support throughout my academic journey.

Contents

Chapter 1

Introduction

1.1 The Challenge of Environmental Pollution

Over the past few decades, the upsurge in environmental pollution by chemical compounds has been driven by industrial processes, agricultural methods, consumerism and various other contributing factors. Although these chemicals are integral for many products and have the potential to improve the comfort of modern society, they can also pose risks and adversely affect both human health and the environment, either acutely or chronically. Toxic substances threaten wildlife but also make air, soil, drinking water and food supply less safe.

Nations worldwide maintain comprehensive chemical regulations[1], however, it is anticipated that global chemicals production will double by 2030 [1]. Moreover, the widespread utilization of chemicals, including their inclusion in consumer goods, is expected to expand further. Even though there are over 275 million known chemical compounds registered by the *Chemical Abstracts Service* [2], merely a tiny fraction of them undergo close monitoring via target analytical approaches and even less is known about their toxicity profiles and negative health effects on organsims. Refer to Table 1.1 for an overview of omnipresent water pollutants.

In light of the rapidly evolving chemical landscape, there is an increasing demand for future-proof, robust measurement and modeling methods. These methods are essential for evaluating the toxicity and exposure of chemicals, facilitating informed risk-based decision-making even when data on hazards and exposures are limited. It is worth noting that the need for adaptable approaches in chemical safety and sustainability efforts must also prioritize cost-efficiency and gain widespread acceptance among regulatory bodies, industry stakeholders, and the general public.

For instance, the EU has introduced the 8th Environment Action Programme, as out-

[1] For instance, REACH, short for Registration, Evaluation, Authorisation, and Restriction of Chemicals, is an EU regulation aimed at improving chemical safety and allocating risk management responsibilities to companies operating in various sectors.

lined in its European Green Deal (citegreendeal), to provide direction for European environmental policy until the year 2030. This program reinforces the EU's ambitious goal of sustainable living within planetary limits, with a forward-looking vision that extends to 2050. Central to this vision is a zero-pollution commitment, encompassing air, water, and soil quality, all while prioritizing the well-being of EU citizens. In 2021, the European Commission introduced a sustainability-focused chemicals strategy [4], which aligns with the EU's zero-pollution ambition. This strategy not only enables the evaluation of the safety and sustainability of both existing and future chemical compounds but also aims to reduce concerning substances, such as *per- and polyfluoroalkyl substances (PFAS)*, through substitution or phasing out wherever feasible. In parallel, the *U.S. Environmental Protection Agency (EPA)* shares a similar scientific consensus and is at the forefront of assessing the potential impacts of chemicals on human health and the environment. Leveraging advanced toxicological and exposure methods, EPA actively promotes risk reduction efforts through its own Chemical Safety for Sustainability National Research Program. This program builds upon the achievements of research initiatives like *ToxCast*[2], *Tox21*[3], and the Endocrine Disruptor Screening Program in the 21st Century (EDSP21), demonstrating a commitment to advancing chemical safety on a global scale.

1.2 The Imperative for Prioritization and Toxicity Assessment

Modern analytical techniques, including *high resolution mass spectrometry (HRMS/MS)*, are gaining significance across various domains such as metabolomics, drug discovery, environmental science and toxicology [5].

In environmental monitoring, the application of nontarget HRMS/MS has notably improved the capacity to detect possibly thousands of contaminants in a single sample. The instrument generates complex spectra that provide information about the masses and fragmentation patterns of compounds present within the sample as illustrated in Figure 1.1. Often, only a minority of these molecules can be definitively identified, while the majority remains unidentified, resulting in their classification into two categories:

- **Identified compounds** are substances for which their chemical structure and properties have been determined and confirmed using additional analytical techniques. These compounds are precisely characterized and can be linked to existing databases or reference spectra, enabling the retrieval of information about their toxicity and other relevant characteristics.
- **Unidentified compounds**, on the other hand, are substances that are detected but lack definitive characterization in terms of their chemical identity, structure, or properties, including its toxicity. Unidentified compounds are observed as peaks or features in these spectra, but their specific chemical attributes remain

[2] https://www.epa.gov/comptox/toxcast
[3] https://tox21.gov/

unknown. Further examination of these compounds is necessary, but it entails a substantial investment of time and resources, emphasizing the importance of prioritization.

When it comes to prioritizing unidentified compounds for further comprhensive testing, the standard approach has been to rely on signal intensity from fragmentation data. However, this approach tends to fall short in delivering an accurate assessment of environmental exposures because the signal intensity may not relate proportionally to the compound's concentration in the sample. Furthermore this approach overlooks the toxicological factors essential for prioritizing compounds with concers related to environmental hazards. As a result, substances with the potential for severe ecological consequences, such as endocrine-disrupting compounds, often go undetected because of their low abundance, even though they exhibit high levels of toxicity. Hence, a pressing need exists for alternative approaches to prioritize unidentified nontarget HRMS/MS signals based on their hazard potential. By incorporating relevant toxicity factors into the equation:

$$\text{Risk} = \text{Hazard} \times \text{Exposure} \tag{1.1}$$

we augment the capacity to make well-informed decisions when evaluating the environmental risk associated with chemicals.

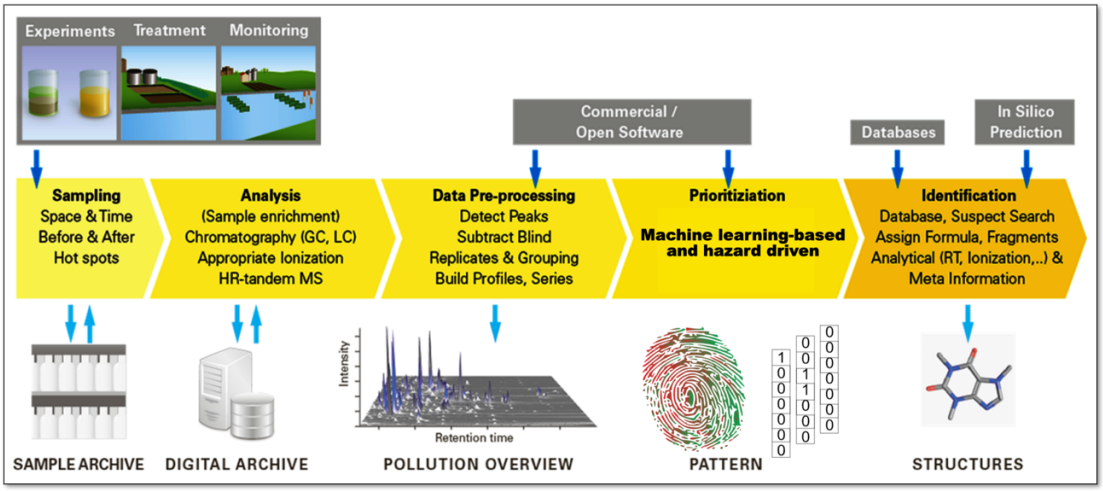
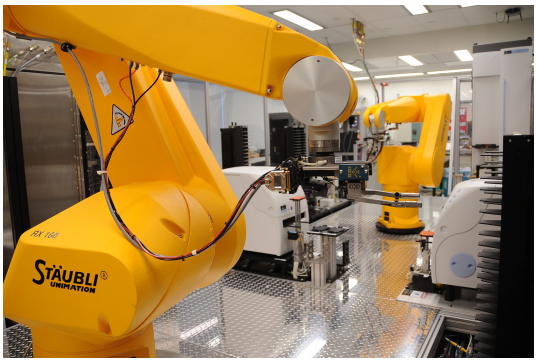


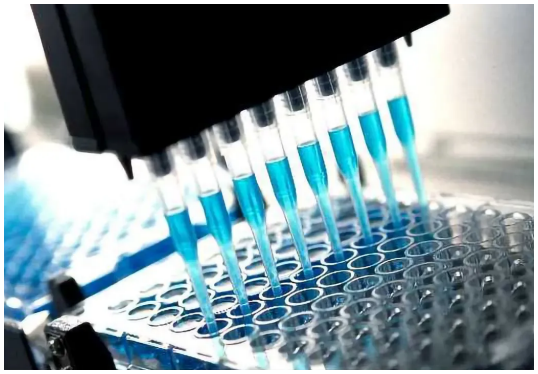
Figure 1.1: Schematic of the workflow used for nontarget HRMS/MS screening of environmental samples, featuring a customized prioritization step. Adapted from Figure 1 in the original source [6].

1.3 Unlocking the Potential of High-Throughput Screening and Machine Learning in Toxicity Prediction

In the past few years, the use of machine learning methods has emerged as a transformative force in the field of *in vitro* toxicology, particularly in the realm of high-throughput toxicity prediction. *High-throughput screening (HTS)* has revolutionized the way toxicity is assessed by allowing thousands of *in vitro* bioassays to be conducted efficiently. This high-throughput approach, coupled with advancements in robotics and automated analysis, has generated large volumes of toxicity data, paving the way for more comprehensive assessments of chemical compounds. Alongside the rise of machine learning, this advancement has facilitated the creation of predictive models, known as *Quantitative structure-activity relationship (QSAR)* models. These models are capable of forecasting bioactivity or compound toxicity based on their physicochemical properties or molecular descriptors [7]. As they are trained on extensive datasets containing comprehensive toxicity information, these models can learn the underlying patterns and relationships between chemical structures and target toxicity. With this capability, they can predict the toxicity of new compounds, even when these substances themselves have not undergone laboratory testing. This approach holds the potential to substantially decrease the time and expenses linked to initial toxicity pre-assessment, and it plays a pivotal role in determining which compounds should undergo more in-depth testing.



(a) A robot arm retrieves assay plates from incubators and places them at compound transfer stations or hands them off to another arm that services liquid dispensers or plate readers. Efforts in the automation, miniaturization and the readout technologies have enabled the growth of HTS. Image obtained from [8].



(b) Modern microtitre assay plates consist of multiples of 96 wells, which are either prepared in the lab or acquired commercially from stock plates. These wells are filled with a dilution solvent, such as *Dimethylsulfoxide (DMSO)*, along with the chemical compounds intended for analysis. Image obtained from [9].

Figure 1.2: High-Throughput Screening (HTS)

1.4 MLinvitroTox: A Novel Approach

In response to the pressing need for a more hazard-driven and comprehensive assessment of environmental contaminants, *Arturi et al.* introduced *MLinvitroTox* [10], an innovative machine learning framework. This framework is part of a broader pipeline named *EXPECTmine*, which incorporates the complementary exposure aspect within the risk assessment process. The primary objective of this thesis is to collaborate with the authors to further enhance and advance this framework. MLinvitroTox leverages molecular fingerprints extracted from fragmentation spectra, marking a significant change in how the toxicity of the myriad unidentified HRMS/MS features is forecasted. MLinvitroTox follows a similar training approach as traditional QSAR models, using supervised classification models trained with molecular fingerprints derived from chemical structures. However, during the application phase, the input to the machine learning model consists of molecular fingerprints generated from experimentally measured*MS2* spectra using *SIRIUS* and *CSI:FingerID* [11]. SIRIUS is a software package for annotating small molecules from nontarget HRMS/MS data, while CSI:FingerID is a machine-learning tool employed by SIRIUS to predict molecular fingerprints from fragmentation spectra.Utilizing streamlined machine learning methodologies, MLinvitroTox forecasts chemical toxicity for a wide range of compounds. This comprehensive analysis covers more than 400 target-specific and 70 cytotoxic endpoints, drawing data from ToxCast/Tox21 datasets. Subsequently, the toxicity predictions generated by the framework are employed to prioritize compounds, with the flexibility to emphasize specific aspects of toxicity profiles tailored to individual preferences.

1.5 Objectives and Significance

The main objective of this thesis is to contribute to the development of an efficient MLinvitroTox framework for predicting compound toxicity across multiple endpoints. The goal is to enhance the integration of MLinvitroTox by creating an automated pipeline in the `Python` programming language. This pipeline is designed to efficiently address the inherent complexities associated with modeling and processing heterogeneous datasets. Here, the emphasis lies in enhancing the curation and filtering of toxicological data and streamlining the process, which begins with raw concentration-response series data and culminates in the generation of the final toxicity predictions. The ultimate output is expected to comprise toxicity fingerprints that encapsulate the predicted toxicity from HRMS/MS environmental samples for the relevant endpoints of interest. These generated toxicity fingerprints will offer crucial insights for the prioritization process, aiding in the identification of the most hazardous compounds present in environmental samples.

One notable contraint of the existing framework lies in its binary *hitcall* when predicting the toxicity of specific endpoints. It categorizes compounds as either toxic or non-toxic without accounting for variations in toxicity severity. In the long term, it is crucial to adopt a more refined approach that can capture the nuanced continuum of toxicity.

This thesis endeavors to overcome this limitation by developing a pipeline capable of forecasting toxicity across numerous endpoints, employing continuous hitcalls.

1.6 Thesis Structure

In the course of progressing through the subsequent chapters, insights will be provided into the materials and methods employed, focusing on the technical intricacies involved in the preparation of ToxCast/Tox21 toxicity data and their transformation into suitable inputs for the machine learning pipeline. This foundational work will establish the basis for the upcoming chapters, which will showcase the potential of MLinvitroTox. Furthermore, the framework's effectiveness is demonstrated through the validation of real-world mass spectral data from *MassBank* [12], and the examination of the implications of this research is carried out.

Origin/Usage	Class	Examples	Related Issues
Industrial Chemicals	Solvents	Tetrachloro-methane	Drinking-water-quality
	Intermediates	Methyl-t-butylether	Drinking-water-quality
	Petrochemicals	BTEX (benzene, toluene, xylene)	Cancer
Industrial Products	Additives	Phthalates	Endocrine disruptors
	Lubricants	PCBs	Biomagnification
	Flame Retardants	PBDEs	
Consumer Products	Detergents	Nonylphenol ethoxylates	Endocrine effects
	Pharmaceuticals	Antibiotics	Bacterial resistance
	Hormones	Ethinyl estradiol	Feminization of fish
Biocides	Pesticides	DDT	Toxic effects and persistent metabolites
	Nonagricultural biocides	Tributyltin	Endocrine effects
Geogenic & Natural Chemicals	Heavy Metals	Lead, cadmium, mercury	Organ damage
	Inorganics	Arsenic, selenium, fluoride	Drinking-water-quality
	Taste and Odor	Geosmin	
	Human Hormones	Estradiol	Feminization of fish
Disinfection & Oxidation	Disinfection by-products	Haloacetic acids, Bromate	Drinking-water-quality
Transformation Products	Metabolites from all above	Metabolites of perfluorinated compounds	Bioaccumulation
		Chloroacetanilide herbicide metabolites	Drinking-water-quality

Table 1.1: Examples of ubiquitous water pollutants. Table 2 adapted from [3].

Chapter 2

Background

This chapter is vital for understanding the following sections of this thesis as it provides some foundational background information in toxicity testing.

2.1 Toxicity Testing: From In Vitro Assays and Molecular Fingerprints to Predictive Models and Beyond

With the ever-growing amount of chemical compounds entering the environment, traditional experimentation methods face limitations concerning cost and time constraints. Additionally, ethical concerns arise regarding the use of animal trials in *in vivo* experiments.

In 2007, the *U.S. National Academy of Sciences* introduced a visionary perspective and published a landmark report, titled as *Toxicity Testing in the 21st Century: Vision and Strategy*. This report promoted a transition from conventional, resource-consuming animal-based *in vivo* tests to efficient high-throughput *in vitro* pathway assays on cells. This transition paved the way for the realm of HTS, where a multitude of *in vitro* bioassays can be executed, complementing and improving chemical screening. This transformation is made possible by advancements in robotics, data processing, and automated analysis. As a result, this synergy has led to the generation of extensive toxicity datasets like ToxCast and Tox21.

HTS datasets, including ToxCast and other sources, have opened the door to promising applications of machine learning in predictive computational toxicology. These predictive models can be developed to screen environmental samples with limited availability of toxicity data, allowing for the prioritization of further testing efforts. Such models often forecast toxicity using QSARs, which are based on descriptors encoding chemical structures like molecular fingerprints. 1D-Molecular fingerprints encode compound molecules as fixed-length binary vectors, denoting the presence (1) or absence (0) of specific substructures or functional groups.

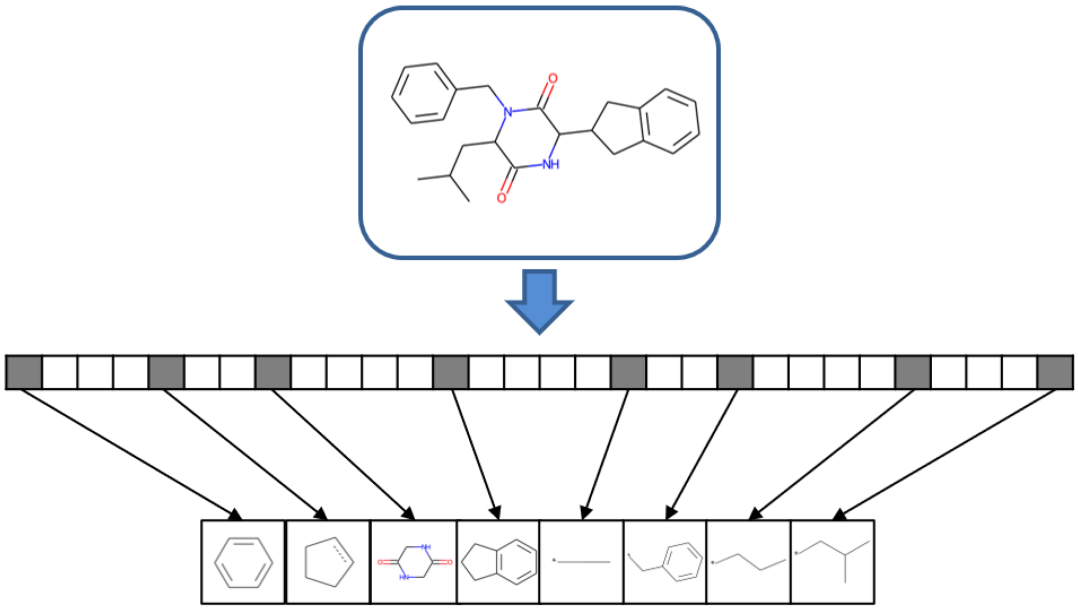


Figure 2.1: Schematic of a molecular fingerprint for a fictional chemical. Each bit position accounts for the presence or absence of a specific structural fragment. Bit positions are set on (set to 1, gray) if the substructure is present in a molecule, or set off (set to 0, white) if it is absent. Figure 1 adapted from [13].

The utilization of molecular fingerprints for *in vitro* toxicity prediction is based on the assumption that molecular toxic effects result from interactions between distinct chemical components and receptors during a *molecular initiating event (MIE)*. On a larger biological scale, the MIE can set a sequential chain of causally linked *key events (KE)* in motion. This occurs at different levels of biological organisation from within cells to potentially culminating in an *adverse outcome pathway (AOP)* at the organ or organism level, as depicted in Figure 2.2. The mechanistic information captured in AOPs reveal how chemicals or other stressors cause harm, offering insights into disrupted biological processes, potential intervention points but also guide regulatory decisions on next generation risk assessment and toxicity testing. The AOP framework is an analytical construct that allows an activity mapping from the presence or absence of certain molecular substructures encoded in chemical descriptors to the target mechanistic toxicity. Finally, when monitoring disruptions in toxicity pathways, physiologically based pharmacokinetic (PBPK) models can be leveraged to extrapolate *in vitro* findings to human blood and tissue concentrations [14].

It is crucial to emphasize that the predictions from HTS bioassays portray molecular toxicity events only at a cellular level, and their translation to adverse outcomes at higher organism levels is not necessarily guaranteed. As the scale shifts from the cellular to the organism level, the confidence in these relationships may decrease.

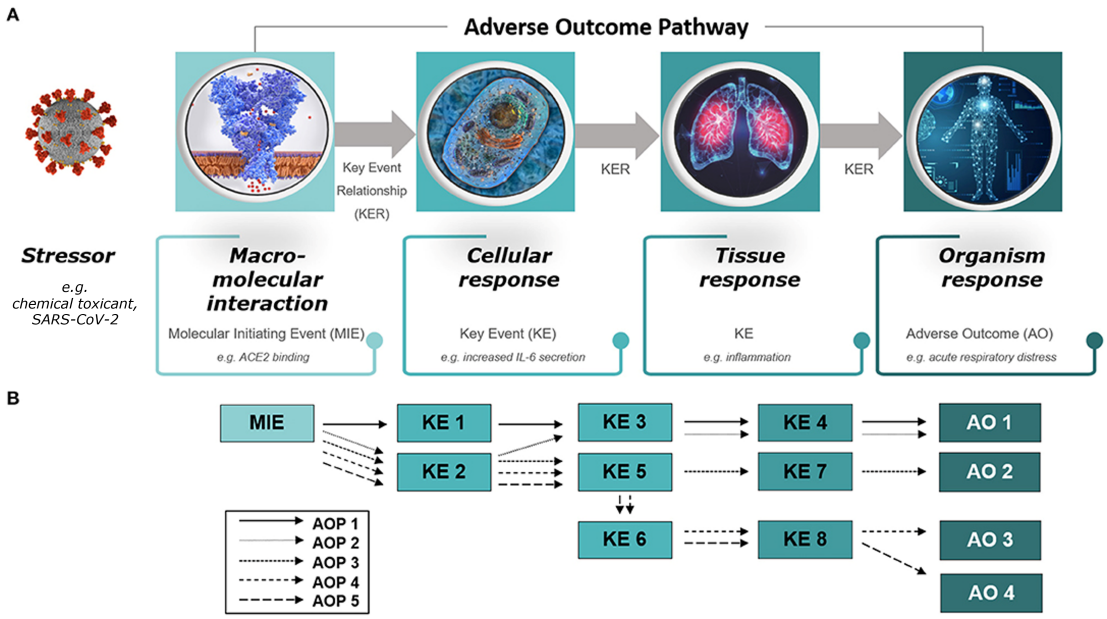


Figure 2.2: Diagram of (A) an adverse outcome pathway (AOP) and (B) an AOP network. (A) An AOP starts with a molecular initiating event (MIE), followed by a series of key events (KEs) on different levels of biological organization (cellular, tissue, organ) and ends with an adverse outcome (AO) in an organism. The stressor is not part of the AOP itself. Figure 1 adapted from [15]

2.2 Chemical Target Toxicity vs. Cytotoxicity

Consider a hypothetical scenario in which a chemical undergoes testing in a bioassay that assesses toxicity by measuring the activation of a *reporter gene* within a cell. The reporter gene encodes a detectable protein, and its activation is triggered by the chemical binding to a specific receptor, the key focus of the assay endpoint. The resulting signal is proportional to the chemical's concentration. While it might seem logical that an increase in chemical concentration would result in higher chemical toxicity, this assumption does not always hold true. At elevated concentrations, the chemical can become *cytotoxic*, causing harm to the cells and ultimately leading to cell death. Consequently, this can lead to a decrease in the activation of the reporter gene and a subsequent reduction in the signal, indicating a decrease in bioactivity. For a visual representation, please refer to Figure 2.3. Considering this situation, chemical toxicity can manifest in various forms, categorizing into two primary groups [16]:

1. **Specific toxicity** is the result of a chemical's interaction and disruption of a specific biomolecular target or pathway, such as a receptor agonist/antagonist effect or enzyme activation/inhibition. This work is primarily concerned with specific toxicity. However, it is essential to recognize that data processing must also take into account the following:

2. **Non-specific toxicity (Cytotoxicity and cell stress)** involve broad disruptions of

the cellular machinery, including reactions with DNA as well as processes like apoptosis, oxidative stress and mitochondrial disturbance. Cell viability can be evaluated either individually or concurrent with the target bioassay endpoint. For instance, one approach involves evaluating the cell viability by determining the proportion of live cells within a population. This is achieved using a fluorescent dye that selectively enters living cells, as it cannot permeate the membranes of deceased cells, resulting in fluorescence intensity directly reflecting cell viability.

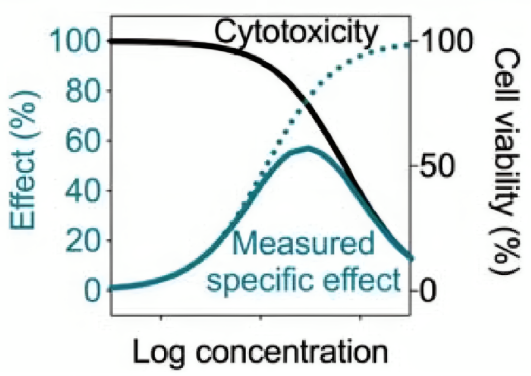


Figure 2.3: Example of a bioassay response with cytotoxicity interference. The dotted line shows the theoretical effect but due to cytotoxicity (black line is cell viability), the measured effect has an inverted U-shape. The measured effect can additionally be confounded by the cytotoxicity burst, where even an exponential curve shape is likely for the gaining part. Figure 7.8 from [17].

An associated phenomenon is referred to as the *cytotoxicity burst* [16], in which the expected specific toxicity interferes with non-specific cellular stress responses that may become overly activated within a critical range of toxicant concentration. As the concentration of the toxicant approaches levels that cause cell death, the signal measuring the supposed specific toxicity of a target assay endpoint becomes mixed with signals from non-specific responses [17]. Compounds that attain an efficacy response xceeding the toxicity threshold within the tested concentration range solely due to these non-specific responses are termed *false positive* hitcalls. This introduces uncertainty about the reliability of reported activity hitcalls, and false positive hitcalls can arise without a comprehensive evaluation of cytotoxicity interference.

The ToxCast pipeline is intentionally designed to minimize *false negative* hitcalls by adopting an inclusive risk assessment approach, ensuring that potentially toxic compounds are not overlooked. Nevertheless, the occurrence of false positive hitcalls can be mitigated through a comparison of potency concentrations between the target assay endpoints and the respective viability or burst assay endpoints that quantify cytotoxic cell loss or cell stress. If the probabilities suggest that the potency concentration of the cytotoxicity assay endpoint is lower than that of the target assay endpoint, previously identified false positive hitcalls can be reevaluated as potential instances of cytotoxicity interference.

Chapter 3

Related work

Similar to MLinvitroTox, MS2Tox [18] represents another machine learning approach within the realm of predicting ecotoxicological hazards for unidentified compounds through nontarget HRMS/MS analysis. Both approaches adopt a common strategy of building their ML models based on molecular fingerprints derived from chemical structure, used to make predictions on environmental samples, utilizing fingerprints from fragmentation spectra calculated by SIRIUS+CSI:FingerID. However, ML2Tox diverges in terms of the toxicity data employed for training and testing, with its focus on toxicity data concerning *in vivo* fish lethal concentrations from CompTox [19]. This is in contrast to MLinvitroTox, which relies on *in vitro* toxicity data from ToxCast/Tox21. Additionally, unlike MLinvitroTox, which exclusively relies on molecular fingerprints and does not utilize any physicochemical properties, MS2Tox incorporates the molecular mass of the compound as an additional feature.

In a systematic investigation using Tox21 data [20], the impact of various modeling approaches on predictive toxicology were explored, with a focus on model performance and explainability trade-offs. The study found that endpoints with higher predictability, characterized by lower data imbalance and larger datasets, performed well regardless of the modeling approach or molecular representation. For less predictable endpoints, simpler models like Linear Regression performed similarly to complex ones, thereby emphasizing the importance of balancing predictability and interpretability. Moreover this study suggests consensus modeling and multi-task learning to enhance predictability and model performance across endpoints. In this thesis, the goal was established to not to overlook simpler models due to their higher interpretability and comparable performance. As recommended, no further explorations were conducted regarding the various molecular representations, and instead, a fixed set of molecular fingerprints was employed as the initial input features, with feature selection being applied to reduce the number of relevant features. Furthermore, a consensus modeling approach was adopted, where the final predictions are obtained by averaging the predictions across assay endpoints sharing the same attributes, including mechanistic and biological target.

Chapter 4

Material and Methods

4.1 Toxicity Data and Processing

4.1.1 ToxCast invitroDB v4.1

The most recent release of the ToxCast's database, referred to as *invitroDBv4.1*, serves as a source of an extensive collection of HTS targeted bioactivity data (100 GB). This database encompasses information on a total of 10 196 compounds, selectively tested across 1485 assay endpoints. Assay endpoints themselves stem from assays, please refer Figure 4.1 for an overview of the assay annotaion structure.

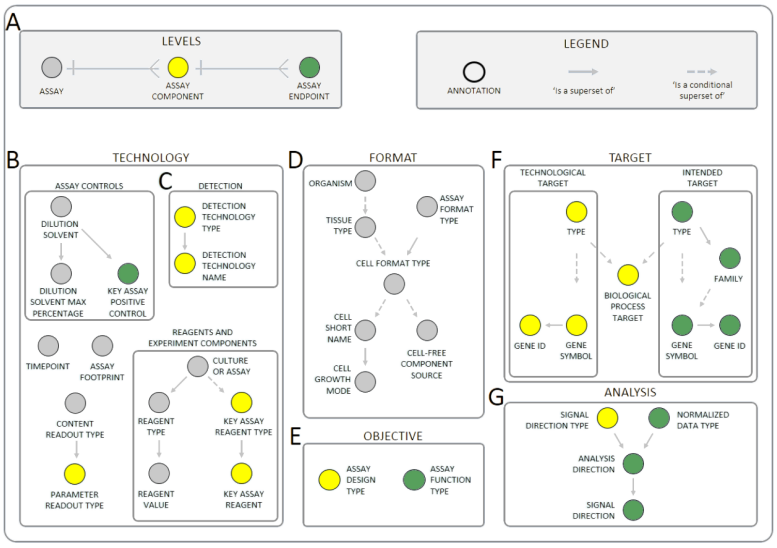


Figure 4.1: The assay annotation structure. Assay endpoints are annotated with (A) assay identification information, (B) design information, (C) target information, and (D) analysis information. Relationships between annotations are either one-to-many or conditional where certain dependencies may not be applicable. Figure obtained from [21].

The assays utilize a range of technologies to assess the impact of chemical compounds on a wide array of biological targets, including individual proteins and cellular processes such as mitochondrial health, developmental processes and nuclear receptor signaling. This resource originates from the collaboration of two prominent institutions: the U.S. EPA through its ToxCast program and the National Institutes of Health (NIH) via the Tox21 initiative. Using data collected from multiple research labs (refer to Table A.1 in the Appendix), this relational database is accessible to the public and can be downloaded[1] by visiting the official ToxCast website.

4.1.2 tcpl v3.0

The *tcpl*[2] package provides a wide range of tools for efficiently managing HTS data. It enables reproducible concentration-response modeling and populates the `MYSQL` database, invitroDBv4.1. The multiple-concentration screening paradigm intends to pinpoint the activity of compounds, while also estimating their efficacy and potency.

In Section 4.2, we introduce *pytcpl* a `Python` reimplementation of the vital components that underpin the entire ToxCast pipeline. It should be noted that these components, as presented in the following, are applicable to both tcpl and pytcpl.

4.1.3 Concentration-Response Series

Each compound c_j tested within an assay endpoint a_i involves the collection of the respective *concentration-response series (CRS)* denoted as CRS_{ij}, showcased Figure 4.2. A CRS is represented as a set of concentration-response pairs:

$$CRS_{i,j} = \{(conc_{1_{i,j}}, resp_{1_{i,j}}), (conc_{2_{i,j}}, resp_{2_{i,j}}), \ldots, (conc_{n_{\text{datapoints}_{i,j}}}, resp_{n_{\text{datapoints}_{i,j}}})\}$$

where $n_{\text{datapoints}_{i,j}}$ varies based on the number of concentrations tested.

In practice, concentrations are often subjected to multiple testing iterations, resulting in the distinct concentration groups with replicates. Table 4.1 presents the key quantities associated with an individual CRS when considering a specific assay endpoint a_i and compound c_i. To visualize the variations in these metrics across the complete set of analyzed CRS in this work, please refer to Figure A.1 in the Appendix.

Concentration-response pairs, along with essential sample information such as well type and assay well-plate indices, can be retrieved by combining tables *mc0*, *mc1*, and *mc3* from invitroDBv4.1, which represent the raw data. A special role is assigned to the control wells, which typically contain untreated samples or samples with a known, non-toxic response. They are used as a baseline to normalize the treated samples and account for any background noise in the assay [22]. The concentrations are transformed to the logarithmic scale in micromolar, while the responses are control well-normalized to either fold-induction or percent-of-control activity:

[1] https://www.epa.gov/chemical-research/exploring-toxcast-data, released on Sept 21, 2023
[2] https://github.com/USEPA/CompTox-ToxCast-tcpl

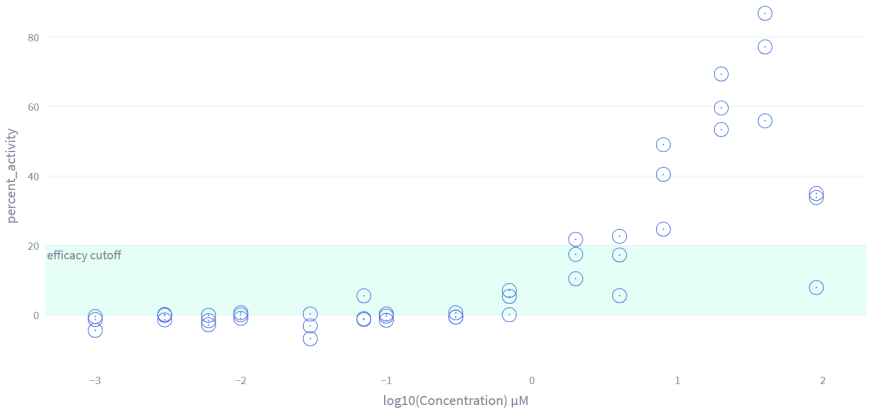


Figure 4.2: The concentration-response series (CRS) belongs to the compound *Diofenolan* (DTXSID2041884), tested in the assay endpoint emphTOX21_ERa_LUC_VM7_Agonist, identified by the assay endpoint ID $aeid = 788$. This particular series comprises a total of $k = 45$ concentration-response pairs and is structured into $n_{conc} = 15$ distinct concentration groups, with each group consisting of $n_{rep} = 3$ replicates.

Table 4.1: Description of Parameters

Quantity	Description
$n_{\text{datapoints}_{i,j}}$	Total number of concentration-response pairs ($\|CRS\|$)
$n_{\text{groups}_{i,j}}$	Number of distinct concentrations tested
$n_{\text{replicates}_{i,j}}$	Number of replicates for each concentration group
$min_{\text{conc}_{i,j}}$	Lowest concentration tested
$max_{\text{conc}_{i,j}}$	Highest concentration tested

1. **Fold Induction**: is a measure used to quantify how much, for instance, gene expression has changed in response to a treatment compared to its baseline level from the control well set. E.g., if a gene is expressed five times higher in a treated sample compared to the control, the fold induction would be 5.
2. **Percent of Control**: is another way to express the relative change in a activity due to a treatment compared to the control.

4.1.4 tcplFit2

TcplFit2[3] is an extension to tcpl, focused on curve-fitting and hit-calling. The package also offers a flexible and robust fitting procedure, allowing for the use of different optimization algorithms and the incorporation of user-defined constraints. This sets it apart from other open-source CSS modeling packages such as *drc* and *mixtox*, as it is explicitly designed for HTS concentration-response data.

[3] https://github.com/USEPA/CompTox-ToxCast-tcplFit2

Table 4.2: tcplfit2 Model Details

Model	Label	Equations[1]
Constant	constant	$f(x) = 0$
Linear	poly1	$f(x) = ax$
Quadratic	poly2	$f(x) = a\left(\frac{x}{b} + \left(\frac{x}{b}\right)^2\right)$
Power	power	$f(x) = ax^p$
Hill	hill	$f(x) = \frac{tp}{1+\left(\frac{ga}{x}\right)^p}$
Gain-Loss	gnls	$f(x) = \frac{tp}{(1+\left(\frac{ga}{x}\right)^p)(1+\left(\frac{x}{la}\right)^q)}$
Exponential 2	exp2	$f(x) = a\left(\exp\left(\frac{x}{b}\right) - 1\right)$
Exponential 3	exp3	$f(x) = a\left(\exp\left(\left(\frac{x}{b}\right)^p\right) - 1\right)$
Exponential 4	exp4	$f(x) = tp\left(1 - 2^{-\frac{x}{ga}}\right)$
Exponential 5	exp5	$f(x) = tp\left(1 - 2^{-\left(\frac{x}{ga}\right)^p}\right)$

[1] Parameters: a: x-scale, b: y-scale p: (gain) power, q: (loss) power, tp: top, ga: gain AC50, la: loss AC50

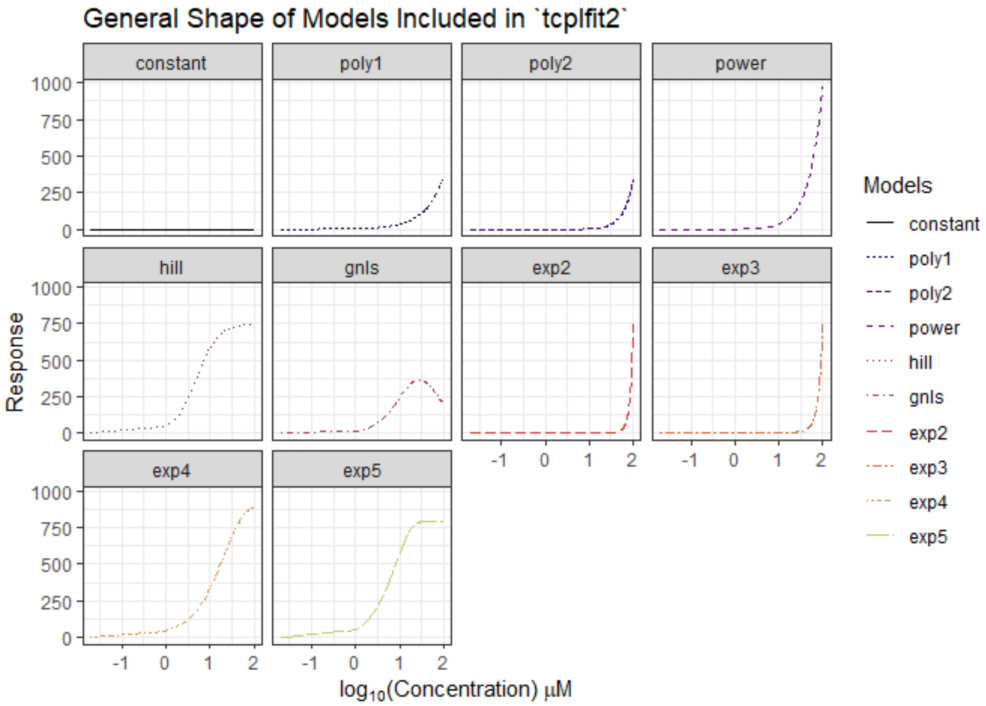


Figure 4.3: Employed curve-fit models in tcpl v3.0 for fitting concentration-response data series through the application of maximum likelihood estimation. Figure obtained from [23].

4.1.5 Curve Fitting

All the curve fit models from tcplFit2, as outlined in Table 4.2 and showcased in Figure 4.3, assume that the normalized observations in the CRS conform to a Student's t-distribution with 4 degrees of freedom [23]. The Student's t-distribution has heavier tails compared to the normal distribution, making it more robust to outlier and eliminates the necessity of removing potential outliers prior to the fitting process. The model fitting algorithm in tcplFit2 employs nonlinear *maximum likelihood estimation (MLE)* to determine the model parameters for all available models.

Consider $t(z, \nu)$ as the Student's t-distribution with ν degrees of freedom, where y_i represents the observed response for the i-th observation, and μ_i is the estimated response for the same observation. The calculation of z_i is as follows: $z_i = y_i - \mu_i \exp(\sigma)$, where σ is the scale term. Then the log-likelihood is: $\sum_{i=1}^{n}[\ln(t(z_i, 4)) - \sigma]$, where n is the number of observations.

The *Akaike Information Criterion (AIC)* is used as measure of goodness of fit, defined by the formula: $AIC = -2\log(L(\hat{\theta}, y)) + 2K$, where $L(\hat{\theta}, y)$ is the likelihood of the model given the data and K is the number of model parameters. The model with the lowest AIC value is chosen as the *winning* model. The winning model is then used to estimate the efficacy and potency of the compound. The potency estimates, also called *point-of-departure (POD)* estimates, are derived from the fitted curve, identifying certain *activity concentrations (AC)* at which the curve first reaches certain response levels. Central POD estimates are depicted graphically in Figure 4.4a.

4.1.6 Hit Calling

The *continuous hitcall* is a measure of the probability that a compound is active, calculated based on the product of the following three probability values [22]:

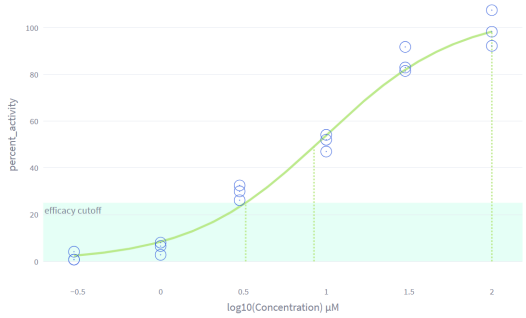
i. that at least one median response is greater than the efficacy cutoff, computed by using the error parameter from the model fit and Student t-distribution to calculate the odds of at least one response exceeding the efficacy cutoff;

ii. that the top of the winning fitted curve is above the cutoff which is the likelihood ratio of the one-sided probability of the efficacy cutoff being exceeded;

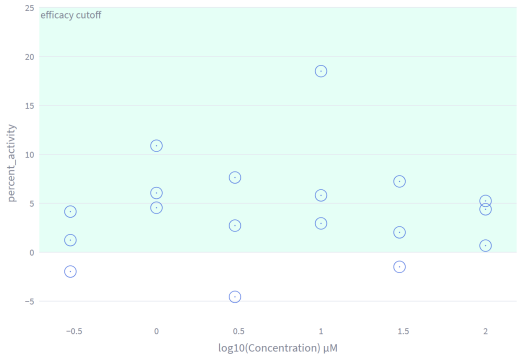
iii. that the winning AIC value is less than that of the constant model:

$$\frac{e^{-\frac{1}{2}AIC_{\text{winning}}}}{e^{-\frac{1}{2}AIC_{\text{winning}}} + e^{-\frac{1}{2}AIC_{\text{cnst}}}} \tag{4.1}$$

In certain instances, compounds underwent multiple tests within a single assay endpoint, leading to their association with multiple CRS. In these exceptional cases, a hitcall is computed for each CRS, and the highest hitcall value is recorded as the compound's ultimate hitcall.



(a) POD estimates for the chemical *Picoxystrobin* (DTXSID9047542) tested in the assay endpoint with $aeid = 753$. The efficacy cutoff is defined at 25 percent-of-control activity. The winning fit model was the Hill function. *ACC*: The AC at the efficacy cutoff is at $3.3\mu M$. *AC50*: The AC at 50% of the maximum response is at $8.4\mu M$. *ACtop*: The AC at the maximum response is at $100\mu M$.



(b) POD estimates are not available for the chemical compound *PharmaGSID_48518* (DTXSID9048518) tested in the same assay endpoint as shown in the left figure. In this case, was unnecessary as no response reached or exceeded 80% of the efficacy cutoff, clearly indicating the inactivity of the compound. In such scenarios, a calculation of POD estimates is not applicable.

Figure 4.4: Presence Matrix: assay endpoint-compound relationship.

4.1.7 Flagging

Finally, after processing, each CRS is categorized into an appropriate fit category based on the level of certainty in the estimated bioactivity. Additionally, cautionary flags are assigned to account for problematic data series or uncertain fits and hits.

4.2 New Toxicity Pipeline Implementation: pytcpl

4.2.1 Introduction

This thesis introduces *pytcpl*[4], a streamlined `Python` repository inspired by the `R` packages *tcpl* and tcplfit2. The package optimizes data storage and generates compressed `Parquet` files of the relevant raw data and metadata from *invitroDBv4.1*. Exclusively utilizing this repository eliminates the need for a complex and extensive database installation, rendering downstream analysis more accessible and efficient. Our package is crafted to accomodate cusomizable processing steps and facilitate interactive data visualization with an own *Curve Surfer*[5]. Furthermore, it enables researchers who prefer `Python` to easily participate in data analysis and exploration, overcoming any limitations associated with using `R` code.

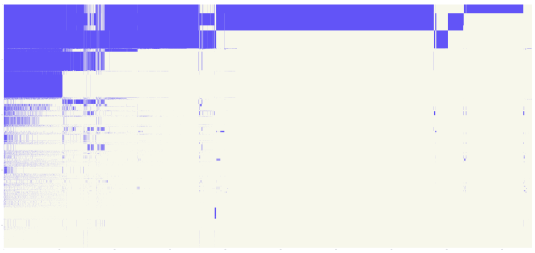
[4] https://github.com/rbBosshard/pytcpl
[5] https://pytcpl.streamlit.app/

The pytcpl pipeline adds an additional setup and wrapup step around the main pipeline:

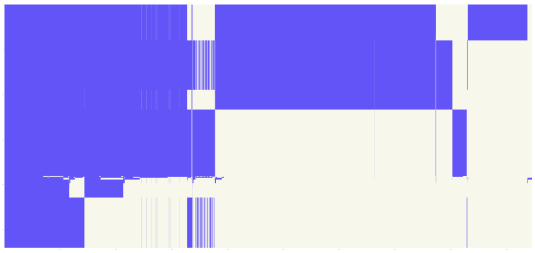
- **Setup**: This step involves user-specified subsetting of assay endpoints, tagging assays with external ICE annotations, enabling workload balancing for distributed processing and generating `Parquet` files from all raw and metadata.
- **Main** (similar to tcpl+tcplFit2): This step involves cutoff determination, curve fitting, hit calling and flagging.
- **Wrapup**: This step involves: ICE post-processing curation, cytotoxicity interference reevaluation and the export of the final results.

4.2.2 Subset assay endpoints

For a better data comprehension, the presence matrix denoted as $P \in 0,1^{m \times n}$ is introduced. In this matrix, rows (indexed by i) represent assay endpoints a_i, and columns (indexed by j) indicate whether testing was performed (1) or not performed (0) for compound c_j in those endpoints. Due to selective compound testing across different assay endpoints, matrix P is sparse. For a visual representation of the presence matrix P covering all assay endpoints and compounds in *invitroDBv4.1*, refer to Figure 4.5a.



(a) The presence matrix P encompasses all assay endpoints and compounds contained within *invitroDBv3.5*, comprising a total of $m = 2205$ assay endpoints and $n = 9541$ compounds. Here, P was structured by sorting it according to the number of compounds associated with each assay endpoint, with compounds arranged in descending order based on their frequency of occurrence. The collective count, where $P_{ij} = 1$, signifies the availability of 3 342 377 concentration-response series (CRS) for subsequent analysisssss.



(b) Modern microtitre assay plates consist of multiples of 96 wells, which are either prepared in the lab or acquired commercially from stock plates. These wells are filled with a dilution solvent, such as *Dimethylsulfoxide (DMSO)*, along with the chemical compounds intended for analysis. Image obtained from [9].

Figure 4.5: Presence Matrix: assay endpoint-compound relationship.

For this thesis, only assay endpoints that have been tested with a minimum of 1000

compounds were exclusively taken into account. This criterion ensures the availability of sufficient data to train a machine learning model with a minimum level of robustness. Please consult Figure 4.5b to view a graphical depiction of the presence matrix P, which includes only the specific considerd subset of assay endpoints. From this point onward, this specific subset will be referred to as the *data* upon which the focus of this thesis will be directed.

In total, $\sum_{i,j} P_{ij} = 1\,372\,225$ concentration-response series are analyzed, encompassing a total of $\sum_{i,j} |S_{ij}| = 48\,861\,036$ concentration-response pairs across all compounds and assay endpoints.

Model	**Label**	**Equations**[1]	**Role in pytcpl**
Exponential 3	exp3	$f(x) = a\left(\exp\left(\left(\frac{x}{b}\right)^p\right) - 1\right)$	Omitted
Gain-Loss 2	gnls2	$f(x) = \frac{tp}{1+\left(\frac{ga}{x}\right)^p} \exp\left(-qx\right)$	New

[1] Parameters: a: x-scale, b: y-scale p: (gain) power, q: (loss) power, tp: top, ga: gain AC50

Table 4.3: tcplfit2 Model Details

4.2.3 Cytotoxicity Interference Evaluation

As introduced in Chapter 2, the measured compound toxicity can be confounded by a non-specific cytotoxicty response. Here, we present the strategy how we reevaluate the reported hitcall of active compouds by the estimated degree of cytotoxicity interference. The cytotoxicity of a compound tested within a certain target assay endpoint can be determined by comparing the activity concentration at the efficacy cutoff (ACC) to the one of its matching viability assay endpoint counterpart. Frequently, assay endpoints have If no counterpart is available in the database, we presented in the following a statistical approach that allows for a cytotoxicity estimate. It uses the median ACC for the compound of interest across a set of assay endpoints dedicted for capturing the cytotoxicity burst. The ACC is assumed to have a Gaussian error distribution. Cytotoxicity in terms of the respective potencies is assumed when: $ACC_{\text{cyto}} \leq ACC_{\text{target}}$. The probability can be expressed as:

$$P(\text{cytotoxic}) = P(ACC_{\text{cyto}} - ACC_{\text{target}} \leq 0) = \Phi\left(\frac{ACC_{\text{cyto}} - ACC_{\text{target}}}{\sqrt{\text{SD}^2_{ACC_{\text{cyto}}} + \text{SD}^2_{ACC_{\text{target}}}}}\right)$$

where Φ is the Gaussian cumulative distribution function. The standard deviations $\text{SD}_{ACC_{\text{cyto}}}$ and $\text{SD}_{ACC_{\text{target}}}$ are unknown but are estimated as $0.3 \log_{10} \mu M$ units [24].

For the statistical approach with the burst assays, $SD_{ACC_{target}}$ can be derived from the median absolute deviation (MAD) of the respective ACC values. Additionally, $P(\text{cytotoxic})$ is muliplied by $\frac{n_{\text{hit}}}{n_{\text{tested}}}$, the ratio of the number where the compound was considerd active divided by the number of cytotoxicity burst assay endpoints where the compound was tested.

Ultitmately, $P(\text{cytotoxic})$ is then mulitplied with the original continuous hitcall of active compounds. The final cytotoxicity-corrected hitcall is then defined as follows: $hitcall_{\text{c}} = hitcall_{\text{original}} * (1 - P(\text{cytotoxic}))$.

4.2.4 Curve Surfer

Figure 4.6 presents the developed *Curve Surfer*, a browser-based application that enables interactive data exploration and visualization of the processed data. The Curve Surfer tool is built using *Streamlit*[6] an open-source Python library that makes it easy to build beautiful custom web-apps for machine learning and data science.

4.3 Machine Learning Pipeline

4.3.1 Preprocessing

Subselecting the columns from the output tables generated by pytcpl: DTXSID identifier and continuous hitcall value. The feature inputs to the machine learning model is a molecular structure represented as fingerprint generated from a SMILES string uniquely determined by the compounds DTXSID identifier. The SMILES string is a linear representation of a compound's molecular structure. The SMILES string is converted to a molecular graph, which is then converted to a feature vector. The feature vector is then used to train a machine learning model. The machine learning model is then used to predict the hitcall value for a given compound. The machine learning pipeline is illustrated in Figure?.

4.3.2 Binary Classification

The goal is to predict whether a compound is active or inactive for a given assay endpoint. This can be formulated as a binary classification problem, where the input consists of the molecular structure fingerprint of the compound, and the output is the binarized hitcall value based on a specific decision threshold. The hitcall value is rendered to a binary variable, where 1 indicates that the compound is active and 0 indicates that the compound is inactive.

4.3.3 Regression

4.3.4 Massbank Validation

[6],

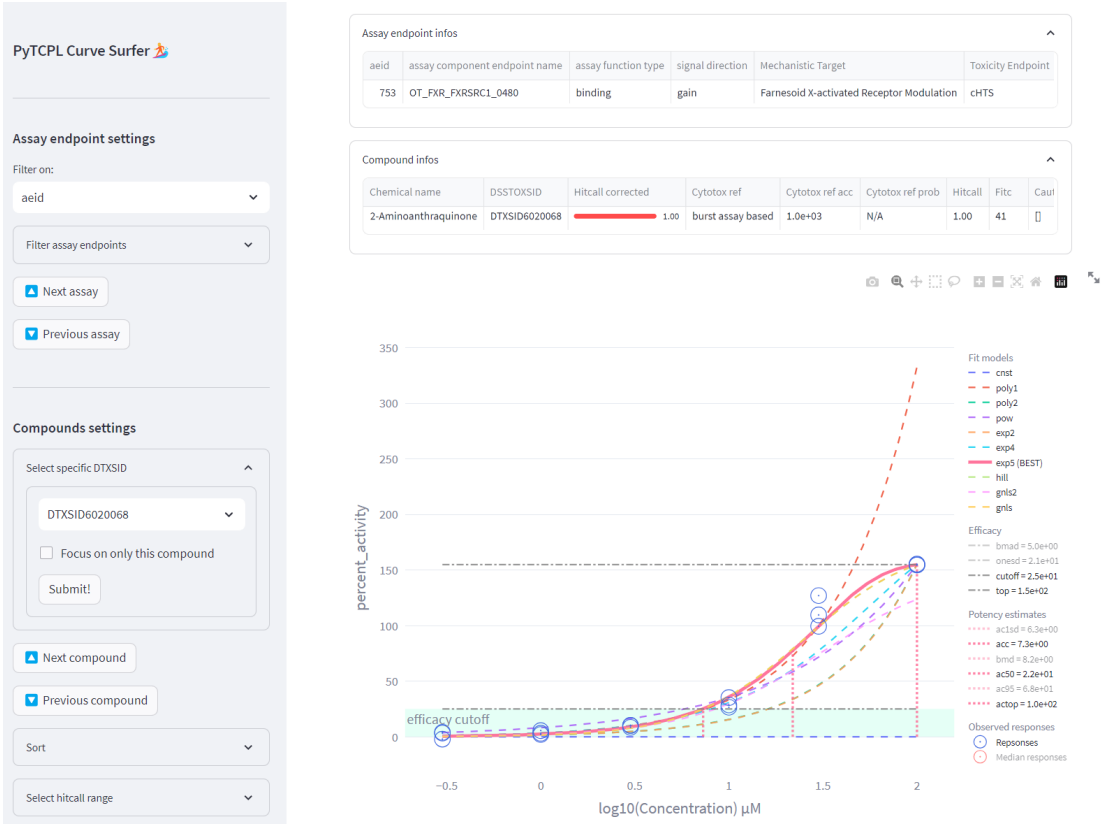


Figure 4.6: Curve Surfer provides the capability to narrow down assay endpoints based on critical annotations, and compounds can be selectively filtered using their DTXSID. Users can navigate through assay endpoints or the compounds within the current assay endpoint. Additionally, compounds can be filtered by their hitcall value or POD estimates using a range slider. Subsequently, Curve Surfer displays comprehensive details for the chosen compound within the opted assay endpoint, showcasing CRS data along with curve fit models and metadata.

Bibliography

[1] U. N. E. Programme, *Global chemicals outlook ii - from legacies to innovative solutions: Implementing the 2030 agenda for sustainable development - synthesis report*, 2019. [Online]. Available: https://wedocs.unep.org/20.500.11822/27651.

[2] C. A. Service, *Chemical abstracts service (cas) is a division of the american chemical society*, Source of chemical information located in Columbus, Ohio, United States, https://www.cas.org/support/documentation/cas-databases, 2023.

[3] R. Schwarzenbach *et al.*, "The challenge of micropollutants in aquatic systems," *Science (New York, N.Y.)*, vol. 313, pp. 1072–7, Sep. 2006. DOI: 10.1126/science.1127291.

[4] E. Commission, "Eu chemicals strategy for sustainability towards a toxic-free environment," 2020, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Chemicals Strategy for Sustainability Towards a Toxic-Free Environment. [Online]. Available: https://environment.ec.europa.eu/strategy/chemicals-strategy_en.

[5] S. Tamara, M. A. den Boer, and A. J. R. Heck, "High-resolution native mass spectrometry," *Chemical Reviews*, vol. 122, no. 8, pp. 7269–7326, 2022, PMID: 34415162. DOI: 10.1021/acs.chemrev.1c00212. eprint: https://doi.org/10.1021/acs.chemrev.1c00212. [Online]. Available: https://doi.org/10.1021/acs.chemrev.1c00212.

[6] J. Hollender, E. L. Schymanski, H. P. Singer, and P. L. Ferguson, "Nontarget screening with high resolution mass spectrometry in the environment: Ready to go?" *Environmental Science & Technology*, vol. 51, no. 20, pp. 11 505–11 512, 2017, PMID: 28877430. DOI: 10.1021/acs.est.7b02184. eprint: https://doi.org/10.1021/acs.est.7b02184. [Online]. Available: https://doi.org/10.1021/acs.est.7b02184.

[7] P. Banerjee, A. O. Eckert, A. K. Schrey, and R. Preissner, "ProTox-II: a webserver for the prediction of toxicity of chemicals," *Nucleic Acids Research*, vol. 46, no. W1, W257–W263, Apr. 2018, ISSN: 0305-1048. DOI: 10.1093/nar/gky318. eprint: https://academic.oup.com/nar/article-pdf/46/W1/W257/25110434/gky318.pdf. [Online]. Available: https://doi.org/10.1093/nar/gky318.

[8] N. H. G. R. I. Maggie Bartlett. "Chemical genomics robot." (2009), [Online]. Available: https://en.wikipedia.org/wiki/High-throughput_screening#/media/File:Chemical_Genomics_Robot.jpg.

[9] J. Rudd. "High throughput screening - accelerating drug discovery efforts." (2017), [Online]. Available: https://www.ddw-online.com/hts-a-strategy-for-drug-discovery-900-200008/.

[10] K. Arturi and J. Hollender, "Machine learning-based hazard-driven prioritization of features in nontarget screening of environmental high-resolution mass spectrometry data," *Environmental Science & Technology*, vol. 0, no. 0, null, 0, PMID: 37279189. DOI: 10.1021/acs.est.3c00304. eprint: https://doi.org/10.1021/acs.est.3c00304. [Online]. Available: https://doi.org/10.1021/acs.est.3c00304.

[11] K. Dührkop *et al.*, "Sirius 4: A rapid tool for turning tandem mass spectra into metabolite structure information," *Nature methods*, vol. 16, no. 4, pp. 299–302, Apr. 2019, ISSN: 1548-7091. DOI: 10.1038/s41592-019-0344-8. [Online]. Available: https://research.aalto.fi/files/32997691/SCI_Duhrkop_Fleischauer_Sirius_4_Turning_tandem.pdf.

[12] *Massbank: High quality mass spectral database*, https://massbank.eu/MassBank/, Accessed: 2023.

[13] T. Janela, K. Takeuchi, and J. Bajorath, "Introducing a chemically intuitive core-substituent fingerprint designed to explore structural requirements for effective similarity searching and machine learning," *Molecules*, vol. 27, no. 7, 2022, ISSN: 1420-3049. DOI: 10.3390/molecules27072331. [Online]. Available: https://www.mdpi.com/1420-3049/27/7/2331.

[14] S. M. Bell *et al.*, "In vitro to in vivo extrapolation for high throughput prioritization and decision making," *Toxicology in Vitro*, vol. 47, pp. 213–227, 2018, ISSN: 0887-2333. DOI: https://doi.org/10.1016/j.tiv.2017.11.016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0887233317303661.

[15] P. Nymark *et al.*, "Systematic organization of covid-19 data supported by the adverse outcome pathway framework," *Frontiers in Public Health*, vol. 9, May 2021. DOI: 10.3389/fpubh.2021.638605.

[16] R. Judson *et al.*, "Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on In Vitro Assay Activity Across a Diverse Chemical and Assay Space," *Toxicological Sciences*, vol. 152, no. 2, pp. 323–339, May 2016, ISSN: 1096-6080. DOI: 10.1093/toxsci/kfw092. eprint: https://academic.oup.com/toxsci/article-pdf/152/2/323/26290632/kfw092.pdf. [Online]. Available: https://doi.org/10.1093/toxsci/kfw092.

[17] B. Escher, P. Neale, and F. Leusch, *Bioanalytical Tools in Water Quality Assessment*. IWA Publishing, Jun. 2021, ISBN: 9781789061987. DOI: 10.2166/9781789061987. eprint: https://iwaponline.com/book-pdf/899726/wio9781789061987.pdf. [Online]. Available: https://doi.org/10.2166/9781789061987.

[18] P. Peets, W.-C. Wang, M. MacLeod, M. Breitholtz, J. W. Martin, and A. Kruve, "Ms2tox machine learning tool for predicting the ecotoxicity of unidentified chemicals in water by nontarget lc-hrms," *Environmental Science & Technology*, vol. 56, no. 22, pp. 15 508–15 517, 2022, PMID: 36269851. DOI: 10.1021/acs.est.2c02536. eprint: https://doi.org/10.1021/acs.est.2c02536. [Online]. Available: https://doi.org/10.1021/acs.est.2c02536.

[19] A. J. Williams *et al.*, "The comptox chemistry dashboard: A community data resource for environmental chemistry," *Journal of Cheminformatics*, vol. 9, no. 1, p. 61, 2017. DOI: 10.1186/s13321-017-0247-6. [Online]. Available: https://doi.org/10.1186/s13321-017-0247-6.

[20] L. Wu, R. Huang, I. V. Tetko, Z. Xia, J. Xu, and W. Tong, "Trade-off predictivity and explainability for machine-learning powered predictive toxicology: An in-depth investigation with tox21 data sets," *Chemical Research in Toxicology*, vol. 34, no. 2, pp. 541–549, 2021, PMID: 33513003. DOI: 10.1021/acs.chemrestox.0c00373. eprint: https://doi.org/10.1021/acs.chemrestox.0c00373. [Online]. Available: https://doi.org/10.1021/acs.chemrestox.0c00373.

[21] J. Phuong *et al.* "Toxcast assay annotation data user guide." (2014), [Online]. Available: https://www.epa.gov/sites/default/files/2015-08/documents/toxcast_assay_annotation_data_users_guide_20141021.pdf.

[22] T. Sheffield, J. Brown, S. Davidson, K. P. Friedman, and R. Judson, "tcplfit2: an R-language general purpose concentration–response modeling package," *Bioinformatics*, vol. 38, no. 4, pp. 1157–1158, Nov. 2021, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab779. eprint: https://academic.oup.com/bioinformatics/article-pdf/38/4/1157/50422999/btab779.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btab779.

[23] C. for Computational Toxicology and U. E. Exposure, *Tcpl v3.0 data processing*, R package vignette for the tcpl package v3.0, CRAN, 2023. [Online]. Available: https://cran.r-project.org/web/packages/tcpl/vignettes/Data_processing.html.

[24] E. D. Watt and R. S. Judson, "Uncertainty quantification in toxcast high throughput screening," *PLOS ONE*, vol. 13, no. 7, pp. 1–23, Jul. 2018. DOI: 10.1371/journal.pone.0196963. [Online]. Available: https://doi.org/10.1371/journal.pone.0196963.

Appendix A

Appendix

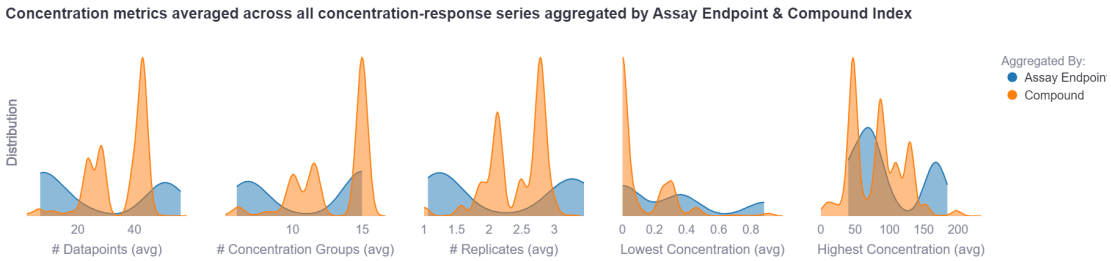


Figure A.1: Concentration metrics averaged across all concentration-response series aggregated by assay endpoint (blue) and compound (orange). E.g., the first chart shows the distribution (blue) on the average number of datapoints across all assay enpoint $a_i \in A$ with $\frac{1}{|C_i|} \sum_j n_{\text{datapoints}_{i,j}}$ and across all compounds $c_j \in C$ with $\frac{1}{|A_j|} \sum_i n_{\text{datapoints}_{i,j}}$. Similarly, the process is repeated for the other metrics: $n_{\text{groups}_{i,j}}$, $n_{\text{replicates}_{i,j}}$, $min_{\text{conc}_{i,j}}$, and $max_{\text{conc}_{i,j}}$.

Table A.1: Assay Source Names and Long Names

assay_source_name	assay_source_long_name
ACEA	ACEA Biosciences
APR	Apredica
ATG	Attagene
BSK	Bioseek
NVS	Novascreen
OT	Odyssey Thera
TOX21	Tox21/NCGC
CEETOX	Ceetox/OpAns
LTEA	LifeTech/Expression Analysis
VALA	VALA Sciences
CLD	CellzDirect
CCTE_PADILLA	CCTE Padilla Lab
TANGUAY	Tanguay Lab
STM	Stemina Biomarker Discovery
ARUNA	ArunA Biomedical
CCTE	CCTE Labs
CCTE_SHAFER	CCTE Shafer Lab
CPHEA_STOKER	CPHEA Stoker and Laws Labs
CCTE_GLTED	CCTE Great Lakes Toxicology and Ecology Division
UPITT	University of Pittsburgh Johnston Lab
UKN	University of Konstanz
ERF	Eurofins
TAMU	Texas A&M University
IUF	Leibniz Research Institute for Environmental Medicine
CCTE_MUNDY	CCTE Mundy Lab
UTOR	University of Toronto, Peng Laboratory